DOCUMENT RESUME

ED 303 998 FL 017 764

AUTHOR de Jong, John H. A. L.
TITLE Le Modele de Rasch: les principes sous-jacents et son application a la validation de tests (The Rasch Model: Underlying Principles and Application to Test Validation).
PUB DATE Mar 88
NOTE 16p.; Paper presented at the Colloquium Meeting of the National Association for the Teaching of French as a Foreign Language (March 11-12, 1988).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Journal Articles (080)
LANGUAGE French
JOURNAL CIT Taaltoetsen: Toegepaste taalwetenschapin artikelen 31; n2 p56-70 1988

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS English (Second Language); French; *Language Tests; Listening Comprehension Tests; Models; Second Language Instruction; *Test Reliability; *Test Theory; *Test Validity
IDENTIFIERS *Rasch Model

ABSTRACT

The one-parameter psychometric model known as the Rasch model is described and examined. The basic principles underlying the model and the concepts of unidimensionality, local stochastic independence, and additivity are explained in non-mathematical terms. The requirements of measurement procedures, the measurement of latent traits, the control on model fit, and the definition of a trait are discussed. It is argued that the Rasch model is particularly appropriate to understand the mutual dependence of test reliability and validity. Examples from French native language and English foreign language listening comprehension tests are used to illustrate the application of the model to a test validation procedure. (Author/MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED303998

TAALTOETSEN

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

toegepaste taalwetenschap in artikelen 31

ABSTRACT in English

THE RASCH MODEL: UNDERLYING PRINCIPLES AND APPLICATION TO TEST VALIDATION

John H.A.L. de Jong

FL 017764

This paper provides an elementary introduction to the one parameter psychometric model known as the Rasch model. It explains the basic principles underlying the model and the concepts of unidimensionality, local stochastic independence, and additivity in non-mathematical terms. The requirements of measurement procedures, the measurement of latent traits, the control on model fit, and the definition of a trait are discussed. It is argued that the Rasch model is particularly appropriate to understand the mutual dependence of test reliability and validity. Examples from foreign language listening comprehension tests are used to illustrate the application of the model to a test validation procedure.

2

jaargang 1988 ISBN. 90.6256.732.0 nummer 2

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
de Jong
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

LE MODELE DE RASCH: LES PRINCIPES SOUS-JACENTS ET SON APPLICATION A LA VALIDATION DE TESTS*)

John H.A.L. de Jong
Instituut voor Toetsontwikkeling (Cito), Arnhem

Cette communication présente une introduction simple au modèle psychométrique connu sous le nom de "modèle de Rasch". Elle traite des rapports entre la fiabilité, la validité et le niveau de difficulté d'un test et présente une méthode pour évaluer la validité. Cette méthode est basée sur la comparaison des résultats d'élèves langue étrangère et d'élèves langue maternelle et est illustrée par des exemples de tests de compréhension auditive d'anglais et de français.

1. *Les qualités requises d'un procédé de mesure*

Imaginons-nous une personne qui a une série de boules de plomb de tailles différentes. Il a aussi une série de boîtes rondes qui ont également des tailles différentes. Il fait des recherches et découvre un rapport très net entre le poids des boules et le nombre de boîtes dans lesquelles elles peuvent être mises: moins une boule pèse, plus de chance elle a de pouvoir être mise dans n'importe quelle boîte; si les boules pèsent plus, il y a moins de boîtes dans lesquelles elles peuvent être casées. Il répète son expérience une fois, deux fois, trente-six fois. Il constate que son test est fiable et il pourra désormais se passer de balances pour évaluer les différences de poids. Mais, une nuit, un voleur de plomb s'introduit dans son laboratoire et lui vole une partie de ses boules de plomb. Le voleur prend les boules les plus

*) Deze bijdrage is een aanpassing van een lezing gepubliceerd in de handelingen van het "Colloque Evaluation dans l'Enseignement des Langues", georganiseerd door de Association Nationale des Enseignants de Français Langue Etrangère (ANEFLE), 11-12 maart 1988,

grosses bien sûr, celles-ci rapportant le plus au marché noir. Pour éviter que son méfait soit découvert, le voleur remplace les boules volées par des boules en plastic soigneusement peintes couleur de plomb. Le lendemain matin notre chercheur revient pour contrôler une dernière fois ses résultats, avant de les publier. Hélas, son test ne suggère plus la même relation entre le poids des boules et le nombre de boîtes dans lesquelles elles peuvent être mises. Ce sont désormais les boules les plus légères, qui ont le moins de chance de trouver une boîte à leur taille. Il a beau répéter l'expérience, il trouve toujours ces résultats contraires. Son test est donc toujours fiable. Il a des problèmes de validité. Un fait reste pourtant certain: si une boule va dans une petite boîte, elle ira aussi dans toutes les boîtes qui sont plus grandes. En réalité son test est un instrument pour mesurer la taille d'objets et non pas leur poids, avant la nuit malheureuse aussi bien qu'après.

La morale de cette histoire: Si l'on veut mesurer quelque chose, il faut d'abord définir l'attribut qu'on va mesurer. Puis, après avoir développé et mis en oeuvre une méthode pour mesurer cet attribut de manière fiable, il faut contrôler la validité, c'est-à- dire: évaluer si l'attribut qu'on a mesuré est vraiment l'attribut qu'on voulait mesurer.

2. *Comment prendre la mesure d'un trait latent*

Dans le cas des tests psychologiques et des épreuves scolaires, nous sommes encore plus dans le noir, car il est souvent impossible d'observer directement l'attribut que nous envisageons mesurer. Il existe pourtant une habitude établie depuis longtemps dans le domaine des épreuves scolaires, celle qui revient à la simple addition du nombre de réponses correctes - ou incorrectes, selon le cas - pour baser ensuite sur ce nombre un jugement de qualité. C'est Georg Rasch (1960) qui a montré qu'en totalisant le nombre de réponses correctes pour obtenir un indice de la compétence ou de l'aptitude d'un candidat, on présuppose en fait que tous les items d'un test mesurent le même trait: que le test ne mesure qu'une seule dimension. Il a prouvé qu'en effet le score total est, comme on dit, une statistique nécessaire et suffisante pour évaluer une aptitude et il a proposé un modèle psychométrique basé sur le présupposé de l'unique dimension dans un test. Ce modèle établit un rapport probabiliste entre le score total d'un candidat et sa réponse à chacun des items dont le test est composé.

Supposons une collection d'items mesurant le même trait et dont on a pu établir l'ordre de difficulté. Un candidat a su répondre a un item de difficulté "x", mais il a échoué devant un item de difficulté "x+1". Si ces deux observations dépendent seulement de l'aptitude du candidat en la matière mesuré par ces deux items, il est probable qu'il saura

bien répondre à un item ayant une difficulté moindre que "x" mais qu'il va sans doute échouer devant chaque item dépassant une difficulté "x+1". En général, il devrait bien répondre à chaque item dans l'intervalle de "x-" à "x" et échouer devant chaque item dans l'intervalle de "x+1" à "x+∞". Plus précisément: dans l'intervalle de "x" à "x+1" la probabilité d'une bonne réponse est de 50%; les items approchant une difficulté de "x+∞ une réponse correcte devient de moins en moins probable et approche enfin 0%; sa chance s'accroît au contraire pour les items dans la direction inverse pour aboutir à une probabilité de 100%.

De la même façon on devrait pouvoir établir une relation entre les candidats essayant de répondre à un item. Si de deux candidats l'un a bien répondu à un item, et l'autre a une compétence plus grande en la matière que l'item prétend mesurer, il est probable que ce dernier donnera également une réponse correcte. La probabilité d'une réponse correcte de la part d'un candidat «plus doué» est plus grande que celle d'un candidat «moins doué», pour chaque item d'un test, à condition que tous les items du test mesurent le même trait.

En comparant items et personnes, on devrait s'attendre à une réponse correcte de la part d'une personne qui a une aptitude plus grande que l'aptitude nécessaire à résoudre l'item, et à une réponse incorrecte de la part d'une personne qui a une aptitude moins grande que l'aptitude nécessaire à résoudre l'item. L'aptitude nécessaire à résoudre un item s'appelle couramment «la difficulté de l'item». De cette façon, il est donc possible de situer la difficulté d'un item et l'aptitude d'une personne sur le même variable.

Résumons:

A condition que tous les items d'un test mesurent le même trait,

- le score total d'un candidat constitue un indice suffisant pour différencier ce candidat de tout candidat qui n'a pas le même score total;
- le nombre total de candidats qui ont répondu correctement à un item suffit à établir la différence entre la difficulté de cet item et les items auxquels un nombre différent de candidats a su trouver la bonne réponse;
- la probabilité d'une réponse correcte de n'importe quel candidat à n'importe quel item dépend entièrement de la différence entre l'aptitude du candidat et la difficulté de l'item: si le candidat a une aptitude qui est exactement égale à la difficulté de l'item, il y a «lutte égale»: une réponse correcte est aussi probable qu'une réponse incorrecte, chacune des deux issues a une probabilité de 50%. Dès que son aptitude est plus grande que la difficulté de l'item, une réponse correcte devient plus probable qu'une réponse incorrecte; si par contre la difficulté de l'item surpasse son aptitude une réponse incorrecte est plus probable.

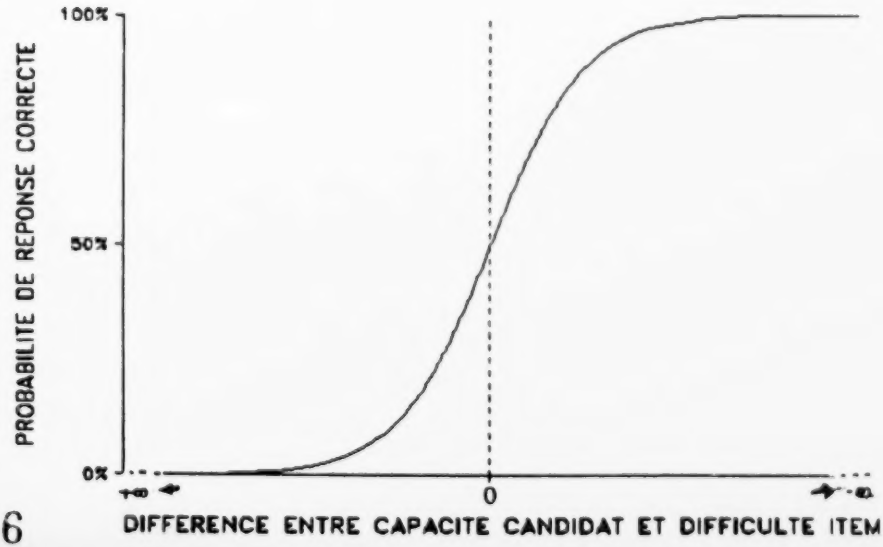
3. *Le contrôle des mesurages*

La relation entre la probabilité d'une réponse correcte d'une part et la différence entre la capacité d'une personne et la difficulté d'un item de l'autre est exprimée mathématiquement dans le modèle élégamment simple de Rasch:

$$\pi(x_{vi}=0,1 \mid \theta_v, \delta_i) = \frac{e^{(\theta_v - \delta_i)}}{1+e^{(\theta_v - \delta_i)}}$$

La probabilité (π) d'observer une réponse (x) de la personne (v) à un item (i) incorrecte (0) ou correcte (1), étant données la capacité (θ) de la personne et la difficulté (δ) de l'item est déterminée par la différence entre la capacité de la personne et la difficulté de l'item. Cette différence est transformée dans une valeur logarithmique népérien (à base e) pour obtenir une fonction montante, continue et couvrant l'intervalle entier entre les extrêmes $-\infty$ et $+\infty$. La fonction de cette relation s'appelle «Courbe caractéristique d'item» et est illustrée dans la Figure 1 ci-dessous.

Figure 1
Courbe caractéristique d'item



Si on accepte ces présupposés, on peut contrôler si les données rassemblées sur un test se conforment en réalité au modèle proposé: après avoir compté le nombre de réponses correctes d'une personne, on estime sa chance de répondre correctement à chacun des items dans le test. Puis on peut évaluer si les réponses données en réalité correspondent aux estimations. Si, par exemple, le modèle évalue la chance pour une personne de répondre correctement à un certain item à, disons, 90%, une réponse incorrecte de la part de cette personne à cet item est extrêmement improbable, et l'observation d'une telle réponse est «suspecte». C'est-à-dire: soit la personne n'a pas travaillé selon ses capacités, soit l'item en question ne mesure pas exactement la même chose que les autres items dans le test. L'attrait de l'adoption d'un modèle psychométrique réside dans la possibilité qu'il nous offre d'évaluer le dégré dans lequel les données rassemblées correspondent aux prédictions du modèle.

4. *La définition de l'attribut*

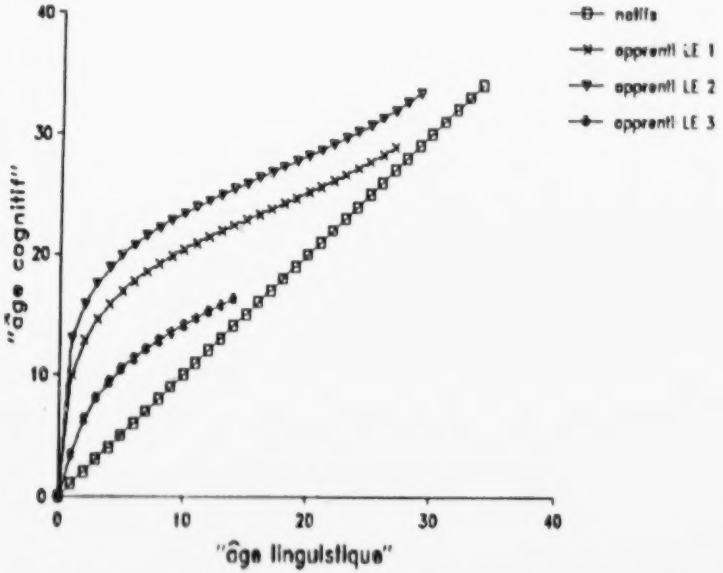
En développant des tests pour évaluer la compétence ou l'aptitude de candidats en langue étrangère, la question est donc de savoir: quel est l'attribut qu'il faut mesurer? Qu'est ce que c'est que la compétence ou l'aptitude en langue étrangère? Faute d'une définition précise toute faite on est obligé de formuler une hypothèse.

Figure 2 présente une hypothèse sur un modèle général d'acquisition de langue. L'axe horizontal représente le développement linguistique d'un sujet à partir de la naissance. D'après notre hypothèse ce développement linguistique ne comporte que les éléments qui sont spécifiques à une certaine langue. L'axe vertical représente le développement cognitif d'un sujet à partir de la naissance. Ce développement cognitif comprend le développement intellectuel dans le sens le plus large: l'acquisition de concepts (cf. e.a. Vygotsky, 1962; Piaget, 1952), des connaissances, et englobe donc également les connaissances linguistiques dites générales, ou universelles et même nombreux éléments et principes spécifiques, mais transférables de la langue à la langue cible (Carton, 1971; Arnaud, 1987). Comme chaque modèle, le modèle proposé est une simplification. Il y a un nombre de dimensions non-représentées, telle que la dimension régionale ou la dimension sociale. Le modèle ne prétend à aucune exactitude en ce qui concerne les distances indiquées sur les axes. Ce qui importe, c'est l'isolation du trait dit «linguistique».

Le développement normal d'un «natif» (personne apprenant sa langue maternelle) est représenté par une ligne droite partant de l'origine et formant un angle de 45° avec les deux axes. Par contre une personne qui débute en langue étrangère aura déjà fait un certain progrès dans

la dimension représentée par l'axe vertical, mais devra se «rattrapper» dans la dimension de l'axe horizontal. Bien sûr le développement cognitif de cette personne ne s'arrête pas pendant qu'elle apprend la langue étrangère. Mais en évaluant ses progrès en langue étrangère, ce développement a peu d'intérêt. Ce qui compte, c'est de mesurer dans la dimension linguistique: dans la seule direction de l'axe horizontal (cf. Chelebourg, 1988).

Figure 2
Modèle hypothétique d'acquisition de langues



Naturellement, la langue, «ça parle de quelque chose», et ce quelque chose doit être compris, c'est une condition sine qua non pour comprendre la langue. Cependant, pour effectuer une évaluation pure de la différence entre deux candidats en langue étrangère, il faut prendre soin à ce que la compréhension ne dépende pas de différences en développement cognitif. Autrement dit: les différences entre les candidats dont on veut évaluer l'aptitude en langue étrangère ne doivent pas dépendre de la compréhension du signifié, mais seulement de la capacité de décoder le signifiant.

Pour évaluer si les scores d'un test reflètent en effet des différences d'aptitude en langue étrangère, on peut comparer les scores de personnes connaissant bien la langue avec ceux de personnes qui doivent

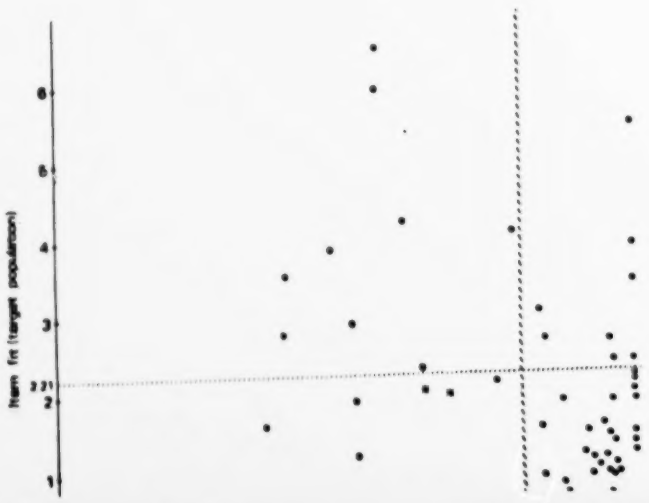
encore l'apprendre. Pour évaluer donc la validité d'un test on peut vérifier si des natifs obtiennent un score plus élevé que des élèves langue étrangère. Bien sûr, les deux groupes doivent avoir atteint environ le même stade de développement cognitif pour assurer que les différences dévoilées par le test dépendent réellement du trait qu'on prétend mesurer.

5. *Correspondance entre l'attribut à mesurer et le trait mesuré*

Dans la Figure 3 chaque point représente un item d'un test expérimental de 59 items de compréhension auditive d'anglais langue étrangère. L'axe de base indique les proportions de réponses correctes données par des Anglais natifs, comparables en âge et en développement aux élèves auxquels le test est destiné. Si 80% de ces natifs répondent correctement à un item, l'item est jugé ne pas présenter de difficultés pour des natifs. Sur l'axe vertical est indiqué l'accord entre les prédictions du modèle de Rasch avec les observations obtenues de l'administration des items à un groupe d'élèves d'anglais langue étrangère. Sur cette axe une valeur dépassant 2.21 indique un manque d'accord et donc une violation de l'assomption d'unidimensionalité: ces items mesurent autre chose que la majorité des items. Les valeurs critiques des deux axes sont marquées avec des lignes pointillées qui coupent le graphique en quatre parties.

Figure 3

Correspondance entre le trait à mesurer et le trait mesuré



Dans la Figure 3 le quadrant en bas à droite comporte le plus grand nombre d'items (35): la plupart des items mesurent un seul trait, qui en plus ne présente pas de problèmes pour des natifs anglais. Dans le quadrant en haut à gauche se trouvent 9 items qui confirment également l'hypothèse que le trait mesuré par le test est réellement la compréhension de l'anglais: car tout en étant trop difficiles pour les natifs ces items ne mesurent pas le même trait que la plupart des items. Pour 44 (35+9) items, soit 75% du nombre total, il y donc correspondance entre d'une part le contrôle de l'unidimensionalité du test effectué par une analyse selon le modèle de Rasch et d'autre part le critère extérieur opérationalisé comme la capacité d'Anglais natifs. Dans une étude antérieure (De Jong 1983) nous avons montré que par un procédé itératif d'analyses, en éliminant chaque fois les items qui dépassent les deux valeurs critiques, il est possible d'augmenter le nombre d'items pour lesquels il y a correspondance entre l'attribut à mesurer et le trait mesuré. Dans l'exemple cité cette correspondance monte à 54 items, soit 92% du nombre total dans le test.

6. *Quelques exemples de mesurages effectués*

Ci-dessous sont présentées deux paires d'items pris dans des essais de tests de compréhension auditive: deux items d'un test de français et deux d'un test d'anglais. Les tests se composaient au total de 42 et de 59 items respectivement. Ces tests ont été administrés à trois groupes de personnes: un groupe d'élèves néerlandais, apprenant le français ou l'anglais selon le cas. Les deux autres groupes étaient des natifs: un groupe correspondant en âge et en formation au groupe néerlandais et un groupe plus jeune et suivant une formation de niveau moins élevé que le groupe néerlandais. Le premier item de chaque paire est un exemple d'un item qui, dans l'analyse selon le modèle de Rasch des résultats des élèves néerlandais, ne paraissait pas mesurer le même trait que la majorité des items dans le test. Le deuxième de chaque paire est au contraire un item qui correspond très bien au trait mesuré par le test. A part le texte enregistré dont on demande la compréhension par les apprenants de la langue cible, et le texte de question à choix multiple destinée à vérifier cette compréhension, est indiqué également le nombre de personnes de chacun des trois groupes qui a choisi chacune des options.

Deux items d'un test de compréhension auditive de français:

14 Texte enregistré:

Est-ce qu'il est toujours utile de faire de la publicité?
Ça, ça, ça dépend tellement du type du produit, du type de marché, du, du dégré de concurrence qui peut exister entre, entre les différents produits. Moi, je sais, justement j'ai du mal euh, essayons de parler, prendre, de prendre des exemples concrets. Je ne sais pas moi, j'aime beaucoup le whisky par exemple. Ben, il y a des bons et des mauvais whiskies. Mais des bons whiskies il y en a beaucoup, il y en a beaucoup. Alors au fond euh, ils ont tous pou.., pour moi, en tant que consommateur, ils ont tous intérêt à se faire connaître... Je crois qu'il y a peut-être une, un, une idée au fond euh, l'idée que de, d'une espèce de produit idéal qui pourrait se passer de publicité pour être euh, pour être consommé et que ça c'est un peu une idée abstraite, ça correspond pas vraiment à une euh, à une réalité, hein.

14 Question à choix multiple: Réponses choisies en % :

Que dit monsieur X de l'utilité de la publicité pour vendre un produit?	Groupe: Néerl. Age:	Nat1 16.5	Nat2 17.5
A Chaque produit profite de la publicité.	A: 39	59	68
B Les produits de qualité supérieure n'ont guère besoin de publicité.	B: 41	5	20
C Ce sont surtout les produits de luxe qui ont besoin de publicité.	C: 20	36	12

6 Texte enregistré:

Vous accepteriez que.. un de vos enfants travaille comme enquêteur ou enquêtrice?
Au porte-à-porte comme ça? Ben, écoutez, je peux vous dire que non seulement j'accepterais, mais que ma... Il est arrivé à ma fille qui, qui travaille avec moi, c'est une de mes collaboratrices... Et ben, c'est pas moi qui l'aie persuadée de le faire, c'est elle qui m'a dit qu'elle souhaitait le faire, donc euh... Et il est arrivé à ma fille de le faire et elle euh, elle a passé plusieurs mois à faire ce genre de,

de, d'enquêtes au porte-à-porte ou bien dans d'autres circonstances, oui... hm. Elle a fait ça pendant, pendant plusieurs mois euh oui, et elle en avait assez, elle en avait marre quoi. Et, et moi ce que je lui avais dit, c'était que c'était pas mal qu'elle fasse cette expérience pour se rendre compte euh.., mais que évidemment ce, ça serait bien si elle arrivait à faire euh, quelque chose d'un, d'un, d'un peu plus euh, d'un peu plus élaborée et d'un peu, d'un peu mieux, quoi.

6 Question à choix multiple: Réponses choisies en % :

La fille de monsieur X a travaillé comme enquêtrice. Qu'en dit-il?	Groupe: Néerl. Nat1 Nat2		
	Age: 17.5	16.5	17.5
A Il a approuvé qu'elle ait voulu faire ce genre de travail.	A: 48	95	96
B Il avait lui-même proposé qu' elle essaye de faire ce genre de travail	B: 34	5	4
C Il lui a d'abord déconseillé de faire ce genre de travail.	C: 18	0	0

L'item nº 14 présente des difficultés de compréhension pour les deux groupes de français qui ont essayé d'y répondre *et*, rappelons-nous, l'item 14 était aussi l'item qui, d'après l'analyse de Rasch, ne mesurait pas le même trait que la majorité des item dans le test. Il y donc une correspondance entre le trait mesuré et le trait à mesurer. En plus l'item nº 14 suggère une distance nette entre les deux groupes de natifs, qui se distinguent par une différence d'âge et de niveau de formation. Apparemment l'item mesure sur une dimension relatif à ces différences. En analysant le texte enregistré et la question proposé on se rend facilement compte où réside la difficulté: il faut suivre exactement le raisonnement de la personne interviewée avec toutes ses hésitations et reprises et non pas prendre trop littéralement ce qu'elle dit au début du fragment. Ce qu'elle dit là sert en fait à gagner du temps, la personne pense à haute voix pour cacher qu'elle ne sait pas tout de suite ce qu'il faut répondre. En plus il faut comprendre que le whisky est pris comme exemple de tous les produits et non pas des produits de luxe. L'auditeur doit au fond lui-même procurer la conclusion générale à partir d'un exemple concret. Toutes ces opérations font appel à un certain dégré de développement cognitif qui, paraît-il, n'est pas évident parmi les élèves néerlandais apprenant le français, ni parmi des élèves français.

pas évident parmi les élèves néerlandais apprenant le français, ni parmi des élèves français.
Bien que l'item nº 6 est également assez difficile pour les apprenants de français, il ne présente guère de difficultés aux deux groupes de natifs. Le fait que l'item nº 6 mesure (selon l'analyse d'après le modèle de Rasch) la même dimension que la majorité des items dans le test, confirme l'hypothèse que cette dimension est en effet une aptitude dans laquelle des natifs excellent, quel que soient leur âge et leur formation. Aussi une analyse de contenu de l'item nº 6 nous montre-t-elle que la compréhension du texte enregistré telle qu'elle est vérifiée dans la question à choix multiple, ne demande qu'une simple opération de paraphrase littérale. Il s'agit de reconnaître «approuver» et «vouloir» dans la question comme synonymes de «accepter» et de «souhaiter» dans le texte enregistré.
Pour augmenter la validité du test d'essai en tant que test de compréhension du français langue étrangère il faudrait, dans la version officielle, éliminer des items comme le nº 14 et garder les items comme le nº 6. Dans une publication antérieure (De Jong & Glas 1987) nous avons montré comment la méthode d'analyse itérative en éloignant d'un test les items ne mesurant pas la dimension voulue peut augmenter la validité de ce test et de quatre autres tests discutés dans cette publication, deux test de français et deux tests d'allemand.
Voici deux autre items, ceux-ci pris d'un test expérimental de compréhension auditive de l'anglais. Par hasard le premier item de cette paire porte aussi le numéro 14.

14 Texte enregistré:

(On parle de moyens électroniques pour payer)

Is it your impression though that, eh, people are in a way reluctant to use these new facilities?
I think one of the problems is, that the credit card is an emotive term in some peoples minds. An enormous number of people who have Barclay cards or Access card, in fact use them for the purchase of goods and there's no credit associated with it at all, and that is going to be true of the new payment systems.

14 Question à choix multiple: Réponses choisies en % :

Some people do not like to buy things on credit.	Groupe:	Néerl.	Nat1	Nat2
	Age	17.5	16.5	17.5
A True	A:	65	49	69
B False	B:	35	51	31

44 Texte enregistré:

Prince Charles appealed for management and unions to reach a better understanding and went on to express his concern that Britain was [---*buzz*---] its overseas competitors.

44 Question à choix multiple: Réponses choisies en % :

Which word could replace the *buzz*? things on credit.	Groupe:	Néerl.	Nat1	Nat2
	Age	17.5	16.5	17.5
A tolerating	A:	70	92	96
B falling behind	B:	30	8	4

Dans cette paire d'items aussi le premier est rejeté par le modèle de Rasch. En accord avec cette observation est le fait que le groupe cadet de natifs anglais montre un score même plus bas que les apprenants. C'est que ici la réponse est encore moins explicite que dans l'exemple français. Dans le fragment on demande si certaines personnes n'hésitent pas à employer les nouveaux moyens. La réponse semble être positive car il est question d'un terme émotionnel, l'écouteur *pourrait* conclure que le public n'aime pas l'idée d'acheter à crédit. L'item demande une gymnastique intellectuelle, qui en soi n'a rien à faire avec la maîtrise d'une langue spécifique, ce qui explique que les deux groupes d'âge égal ont à peu près la même probabilité de trouver la bonne réponse, quelle que soit leur langue maternelle. Le groupe plus jeune a nettement moins de chances.

L'item nº 44 par contre ne présente aucune difficulté pour les deux groupes de natifs et les différencie bien du groupe d'apprenants. De ce test aussi il faut donc éliminer les items comme le numéro 14, mais garder des items comme le numéro 44. (L'analyse de ce test est plus amplement discutée dans De Jong 1984a et 1984b.)

Dans l'introduction de cette communication nous avons expliqué com-

ment un test fiable peut présenter des problèmes de validité. Il faudrait en fait préciser: il n'y pas de fiabilité sans validité, car si un procédé de mesurage se montre fiable il faut qu'il y ait un rappport entre les résultats du mesurage et quelque attribut, manifeste ou latent, des objets mesurés. Le problème est plutôt que souvent on ne sait pas quel attribut on a mesuré, c'est-à dire: à quoi le test est valide. Inversement il n'y a pas non plus de validité sans fiabilité, car comment pourrait-on prétendre qu'on a vraiment mesuré un objet, si à chaque nouvelle prise de mesure, la mesure change? Le grand avantage du modèle de Rasch est justement que les présupposés du modèle partent de cette indivisibilité et forcent le constructeur de test à ce rendre compte que la difficulté d'un item, d'un test doit s'expliquer en termes de l'attribut que l'on veut mesurer. Cette communication a montré, je l'espère, comment on peut dévoiler le caractère du trait latent en se servant du modèle de Rasch pour établir un rapport entre la dimension dans laquelle le test mesure et l'attribut que l'on envisage de mesurer.

Références

Arnaud, P.J.L., 1987. La L1 et l'acquisition de vocabulaire d'une L2. Dans: *Enseignement/apprentissage du lexique en français langue étrangère; Actes du Colloque ANEFLE*, 9-35.

Carton, A.S., 1971. Inferencing: a process in using and learning language. Dans: Pimsleur, P. and Quinn, T. (Eds.), *The psychology of second language learning*. Cambridge University Press, Cambridge.

Chelebourg, C. (1988). Langue et littérature: les problèmes de l'évaluation aux degrés 2 et 3? Dans: A.Edelman et M. Erman (Eds.), *Actes du Colloque ANEFLE*. Dijon.

De Jong, J.H.A.L., 1983. Focusing in on a latent trait: An attempt at construct validation by means of the Rasch model. Dans: Van Weeren, J. (Ed.), *Practice and Problems in language testing 5; non-classical test-theory: Final examinations in secondary schools*. Cito, Arnhem.

De Jong, J.H.A.L., 1984a. Listening, a single trait in first and second language learning. *Toegepaste Taalwetenschap in Artikelen, 20*, 66-79.

De Jong, J.H.A.L., 1984b. Testing foreign language listening comprehension. *Language Testing, 1*, 97-100.

De Jong, J.H.A.L. & Glas, C.W., 1987. Validation of listening comprehension tests using Item Response Theory. *Language Testing, 4*, 170-194.

Piaget, J., 1952. ***The origin of intelligence in children***, International University Press, New York.

Rasch, G., 1960. ***Probabilistic models for some intelligence and attainment tests***. Danmarks Paedagogiske Institut, Kopenhagen.

Vygotsky, L.S., 1962. ***Thought and language***. The M.I.T. Press, Cambridge Mass. (traduction; texte russe original: 1934).